Data Mining - Introduction

Lecturer: JERZY STEFANOWSKI
Institute of Computing Science
Poznań University of Technology
Poznań, Poland

Doctoral School , Catania-Troina, April, 2008

Who … and where …

- Poznań University of Technology, Poland
- Faculty of Computer Science and Management.
- Institute of Computing Sciences,
- **IDSS** = Laboratory of Intelligent Decision Support Systems → http://idss.cs.put.poznan.pl
  - MCDA, Optimization, Evolutionary Computation, Rough and Fuzzy Sets, Machine Learning &Data Mining, Computer Vision.

My research interests

- induction of various types of decision rules from learning examples,
- multiple classification systems and multi-startegy learning,
- learning from imbalance data,
- rough set theory and rough set based learning systems,
- attribute selection for classification systems,
- incomplete data,
- descriptive clustering of text documents and Web pages,
- Web and Text-mining,
- medical data analysis.

Course mission

- Data mining?
  - Extraction of useful information patterns from data.
  - Huge or complex data challenge (DB community).
    - Trend to data warehouses but also flat table files.
  - More than typical data analysis or classical decision support!
  - Still „young” field, although quite „fashionable”.
  - Teaching materials and course book …

Course mission - 2

- Thus, our aims:
  - To give you insight in this field, typical methods, examples of applications.
  - Rather focus on algorithms and methodology aspects, not so much on dealing with massive data.
  - Choice of methods $\rightarrow$ relation to algorithmic decision theory and MCDA.
  - Comments to available software (WEKA, etc.)

Course information

- The course is divided into few parts:
  - Introduction
  - Symbolic methods
    - Decision Trees
    - Decision Rules
  - Non-symbolic methods
  - Advanced topics (multiple classifiers, imbalance)
  - KDD Process and summary

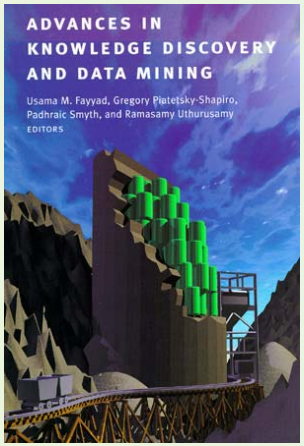
Acknowledgements:

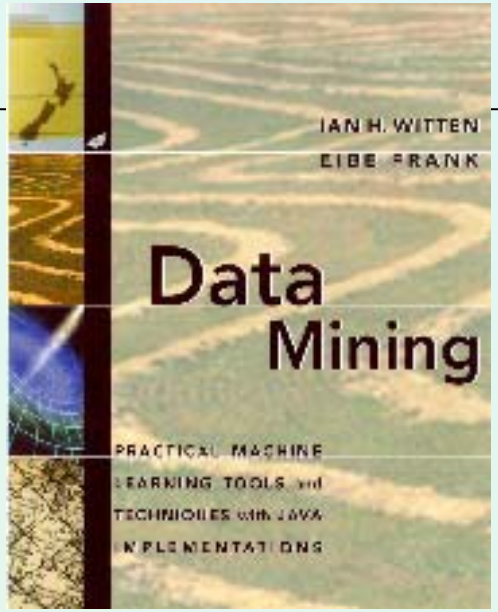
- Many of the slides are based on my own University course:
  - Data mining and Machine Learning (PUT CS; M.Sc. Course) http://www-idss.cs.put.poznan.pl/~stefan
- Some slides are based on ideas borrowed from:
  - WEKA teaching materials (Witten & Frank Waikato University; Morgan Kaufmann)
  - Gregory Piatetsky – Shapiro: Data mining course.
  - Jiawei Han: Knowledge discovery in databases.
  - T.Mitchell and P.Flach courses on ML (see their WWW).
- Other course books – see the next slide

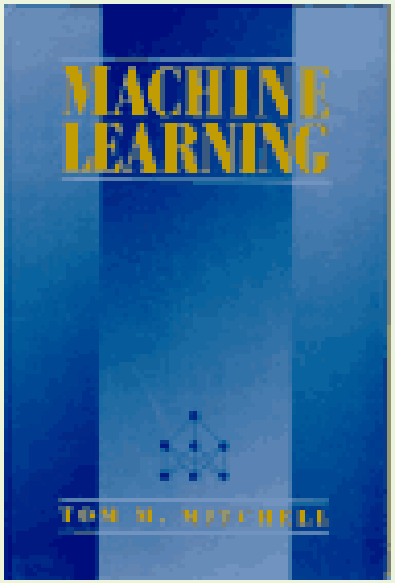
Background literature

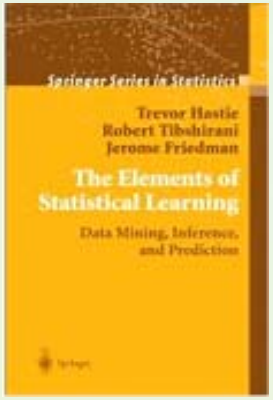
- Han Jiawei and Kamber M. Data mining: Concepts and techniques, Morgan Kaufmann, 2001.
- Hand D., Mannila H., Smyth P. Principles of Data Mining, MIT Press, 2001.
- Mitchell T.M., Machine Learning, McGrawHill, 1997.
- Witten I., Eibe Frank, Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 1999.
- Kononenko I., Kukar M., Machine Learning and Data Mining: Introduction to Priniciples and Algorithms. Horwood Pub, 2007.
- Maimon O., Rokach L., The data mining and knowledge discovery Handbook, Springer 2005.
- Krawiec K, Stefanowski J., Uczenie maszynowe i sieci neuronowe (Machine Learning and Neural Networks), PP Press, 2003.

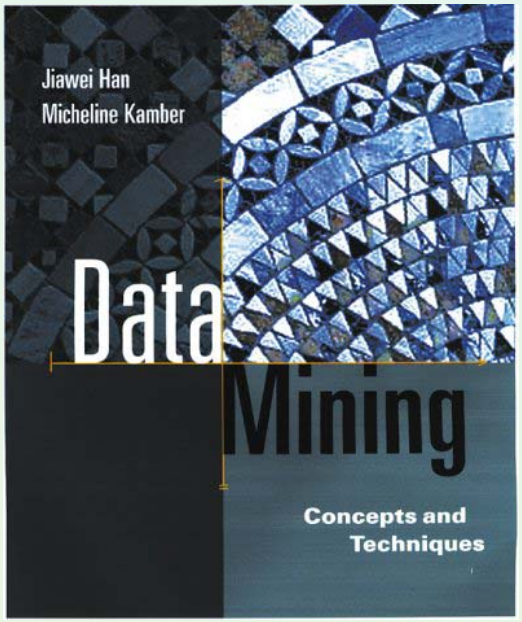
Background literature











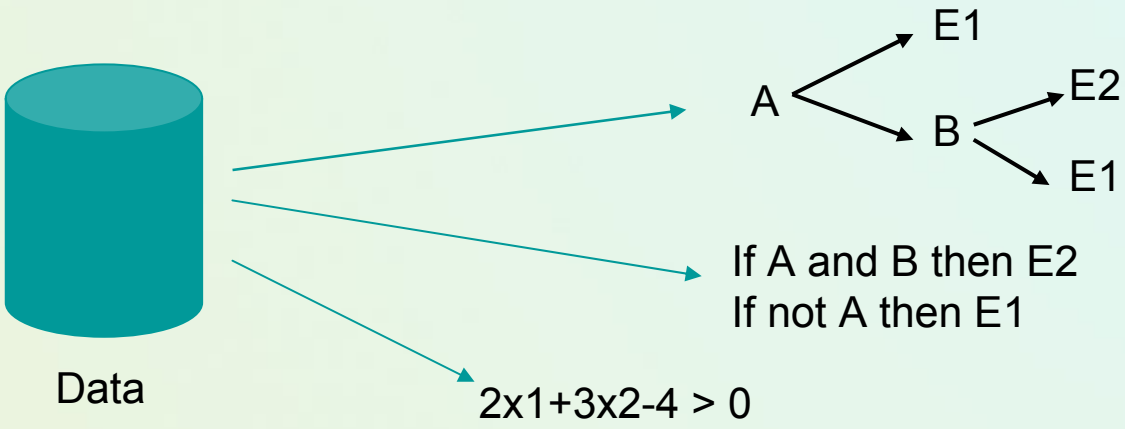
Lecture 1 a.

Data Mining: Introduction

L1. Introduction to Data Mining: Outline

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Basic tasks and methods.
- Examples of applications.
- WEKA – software framework for this course.

Data mining: what is it?



- Data mining is
  - Extraction of useful patterns from data sources, e.g., databases, texts, web, images.
- Patterns (knowledge representation) must be:
  - Valid, novel, potentially useful, understandable to the users.

What is data mining? More …

- Data mining is the analysis of data for relationships that have not previously been discovered or known.
- A term coined for a new discipline lying at the interface of database technology, machine learning, pattern recognition, statistics and visualization.
- The key element in much more elaborate process called „**Knowledge Discovery in Databases**”.
- The efficient extraction of previously unknown patterns in very large data bases.

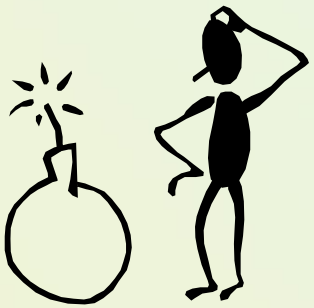
Motivations - data explosion problem

- Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories.
- More data is generated:
  - Bank, telecom, other business transactions ...
  - Scientific data: astronomy, biology, etc
  - Web, text, and e-commerce
- Very little data will ever be looked at by a human!
- We are drowning in data, but starving for knowledge!
- Knowledge Discovery is **NEEDED** to make sense and use of data.

We are Data Rich but Information Poor

Databases are too big

Data Mining can help discover knowledge nuggets



„Terrorbytes”

Knowledge Discovery Definition

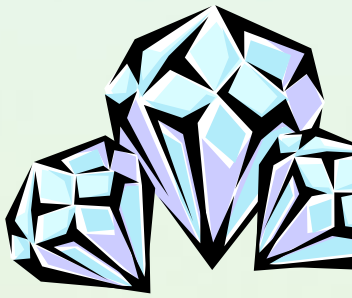
- Knowledge Discovery in Data is the

  non-trivial process of identifying

  - *valid*
  - *novel*
  - potentially *useful*
  - and ultimately *understandable patterns* in data.

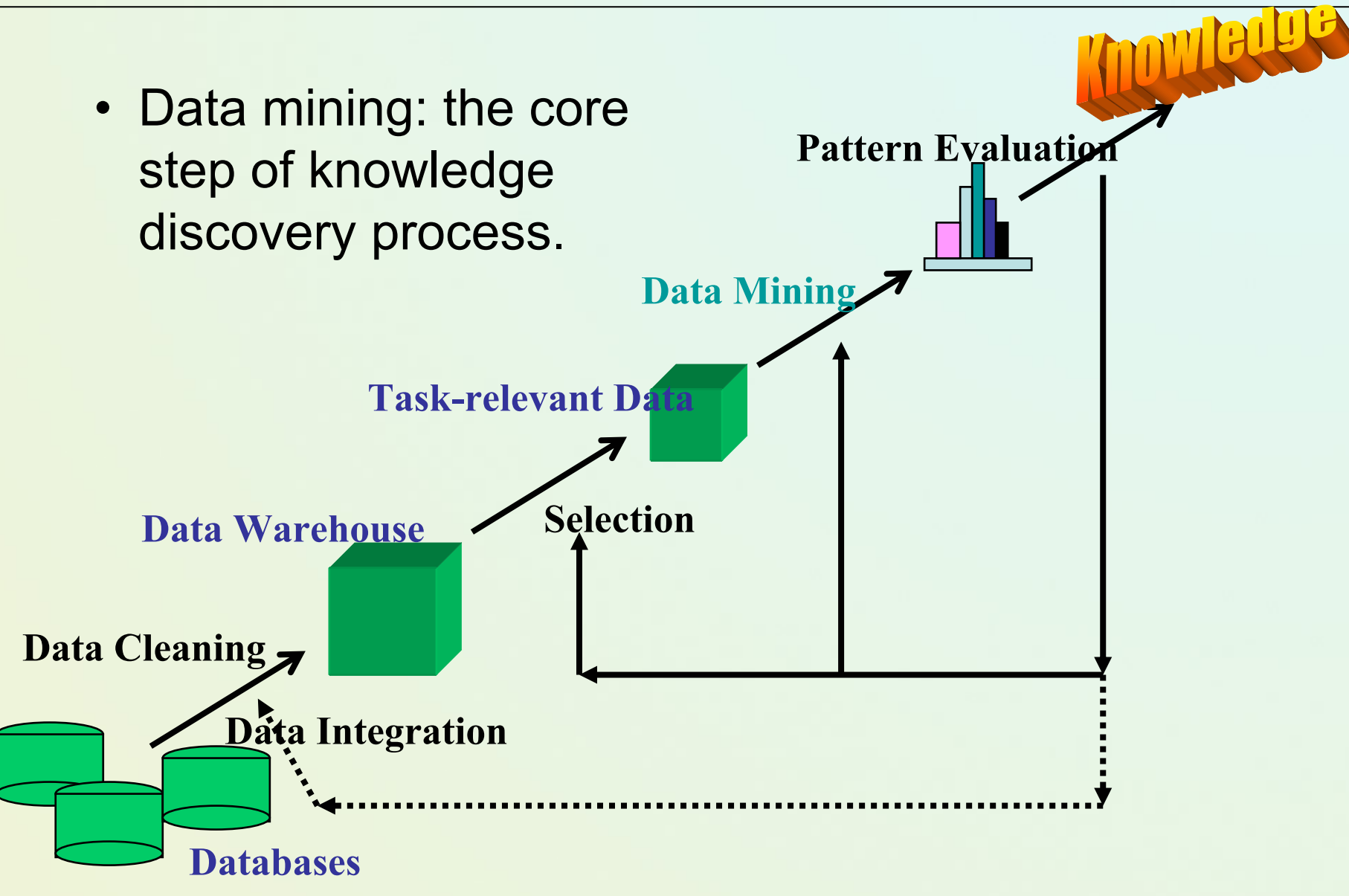
from *Advances in Knowledge Discovery and Data Mining,* Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996.

The name first used by AI, Machine Learning Community in 1989 Workshop at AAAI Conference.



Data Mining as a step in A KDD Process

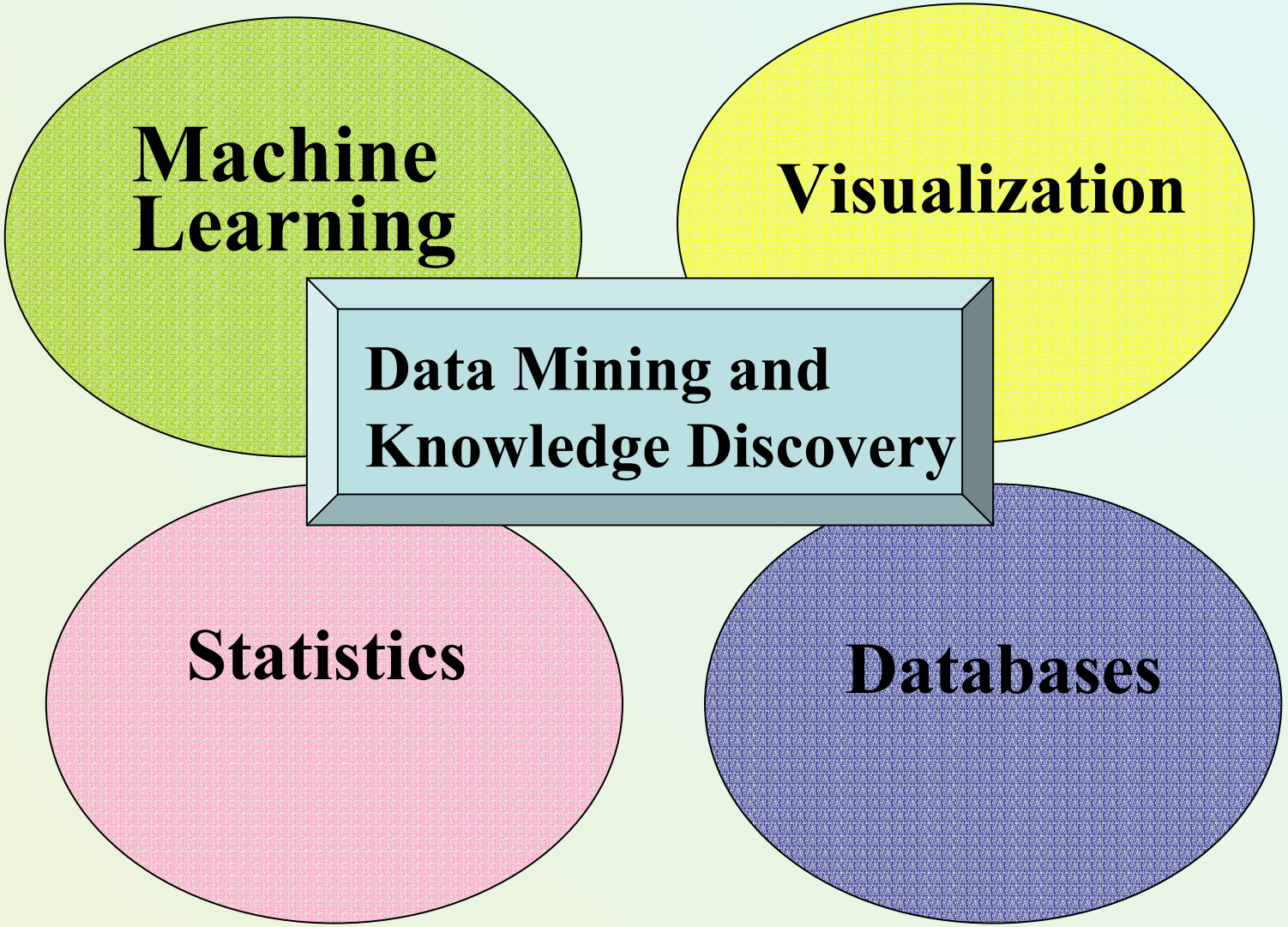
- Data mining: the core step of knowledge discovery process.



Steps of a KDD Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and projection:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Interpretation: analysis of results.
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Related Fields



Data mining and related disciplines

What is not data mining?

- Another statistical approach shifted into a new context!
- It is not only machine learning!
- Moreover, it is not
  - (Data base) query processing,
  - Expert systems software.

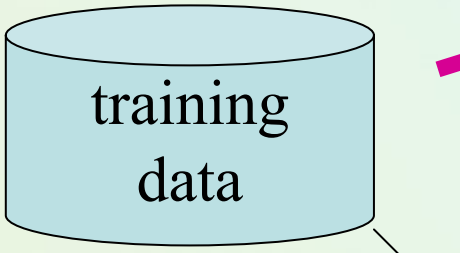
Data Mining: On What Kind of Data?

- **Attribute-value tables**
- Multi-relational data / first order predicate calculus
- Structured data (graphs, workflows, ontologies, …)
- Sequence data bases
- Other advanced data repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - WWW resources
  - …

What can be discovered?

- Symbolic knowledge representations
  - Decision trees
  - Rules
  - Relations / Logic forms (ILP)
  - Attribute generalizations
  - Probability distributions
  - Conceptual clusters and taxonomies
- Sub-symbolic
  - Artificial neural networks
  - Instance based learners
  - Functions / equations
- Others, …

Decision rules



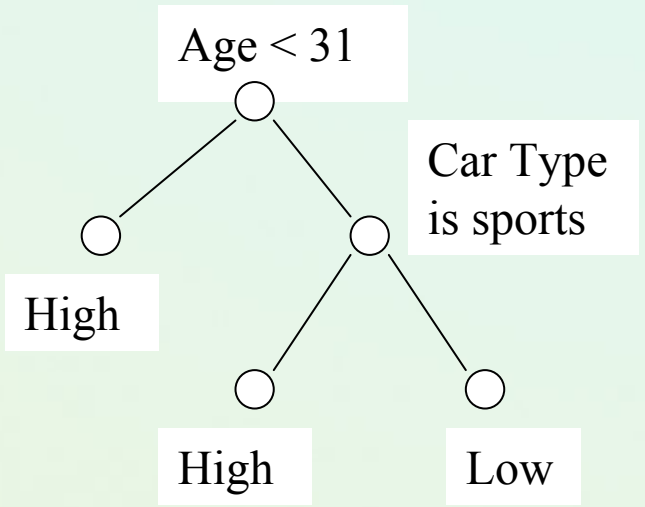
Age	Car Type	Risk
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low

Classification / Learning algorithm

if age < 31
or Car Type =Sports
then Risk = High

Decision trees

Age	Car Type	Risk
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low



Association rules

- Transaction data
- Market basket analysis

TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

- {Cheese, Milk} $\rightarrow$ Bread [sup=5%, conf=80%]
- Association rule:
  „80% of customers who buy *cheese* and *milk* also buy *bread* and 5% of customers buy all these products together”

Sequential Patterns

- Sequential pattern mining is the extracting of frequently occurring patterns related to time or other sequences.
- A sequential rule: $A \rightarrow B \rightarrow C$, says that event A will be followed by event B and this by event C with a certain confidence
- An example:

  “A customer who bought a TV three months ago is likely to order a new DVD player within one month”

Numeric prediction

- Example: 209 different computer configurations

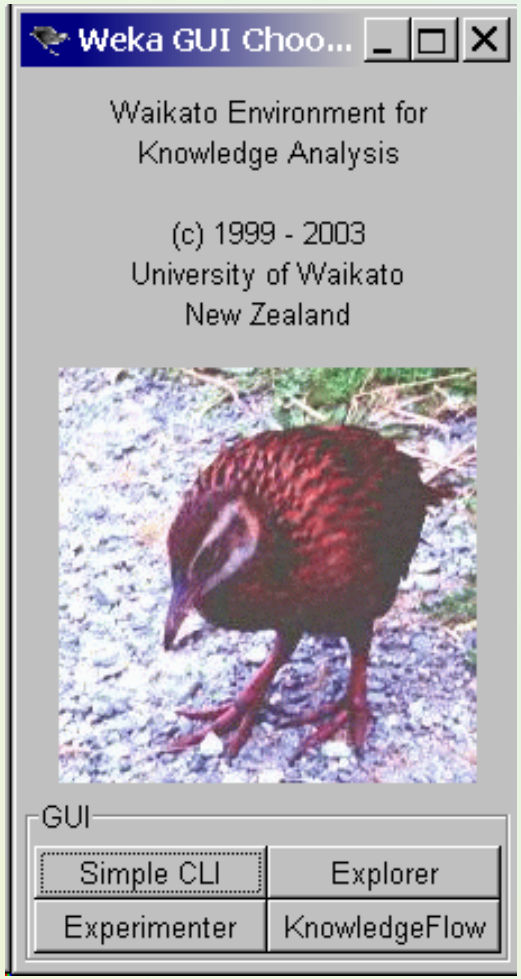
	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
…							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

```
PRP =  -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
       + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

Major Data Mining Tasks

- **Classification:** predicting an instance class on the basis of its description.
- **Prediction**: predicting a continuous value.
- **Clustering:** finding similarity groups in data.
- **Associations:** e.g. A & B & C occur frequently.
  - Sequence analysis
- **Summarization:** describing a group.
- **Visualization:** graphical methods to show patterns in data.
- **Deviation or Anomaly Detection**: finding important changes or anomalies
- …

Weka – software for data mining



- Waikato Environment for Knowledge Analysis (WEKA); developed by the Department of Computer Science, University of Waikato, New Zealand
- Data mining / Machine learning software written in Java (distributed under the GNU Public License)
- Used for research, education, and applications

http://www.cs.waikato.ac.nz/ml/weka/

- Ian Witten, Eibe Frank

WEKA - Tools

- Pre-processing Filters
- Attribute selection
- Classification/Regression
- Clustering
- Association discovery
- Visualization

Why Data Mining? -- Potential Applications

- Database analysis and decision support
  - Market and customer analysis; management
    - target marketing and advertising, customer relation management, market basket analysis, cross selling, market segmentation.
  - Risk analysis and management
    - Forecasting, customer changes, quality control, loan approval.
  - Diagnostics (e.g. technical conditions of objects)
  - Fraud detection
- Other Applications:
  - Text mining (news group, email, documents) and Web analysis; Search engines; Adaptive Web servers and e-commerce systems

Problems Suitable for Data-Mining

- Require knowledge-based decisions.
- Have a changing environment.
- Have accessible, sufficient, and relevant data
- Data are difficult or complex.
- Problems are not trivial, cannot be solved „manually” by humans.
- Provides high payoff for the right decisions!

Privacy considerations important if personal data is involved!

Lecture 1b.

Mining Classification Knowledge: Introduction

Input: concepts-classes, instances, attributes

- In general, the input takes the form of:
  concepts-classes, instances, attributes.
  - Target concepts (classes): kinds of things that can be discovered
    - Aim: intelligible and operational concept description.
  - Instances: the individual, independent examples of a concept
    - Note: more complicated forms of input are possible.
  - Attributes: (features) measuring aspects of an instance
    - We will focus mainly on nominal and numeric ones.

The weather problem (Quinaln's play sport)

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Given past data, Can you come up with the rules for Play/Not Play ?

What is the game?

Training (Learning) Set

- Decision table $(U, A \cup \{y\})$, where U - set of examples, A - set of input attributes $\{a_1, a_2, \ldots, a_n\}$ and y – an output attribute;
- y – either decision/class attribute defined a classification $\{K_1, K_2, \ldots, K_k\}$ or a target numeric attribute (prediction / regression).
- **x** - example described by m attributes
- A training set S = $(\langle x_1, y_1\rangle, \langle x_2, y_2\rangle, \ldots, \langle x_n, y_n\rangle)$
- Aim: to discover an unknown function f : $y = f(\mathbf{x})$

Classification vs. Prediction

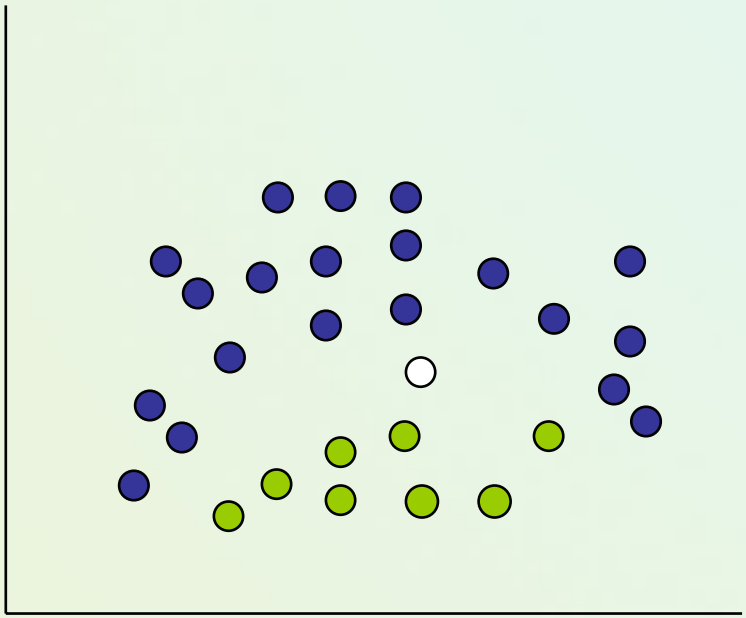
- Classification:
  - predicts categorical class labels (discrete or nominal)
- Prediction:
  - models continuous-valued functions, i.e., predicts unknown or missing values (regression)
- Typical Applications
  - credit approval
  - target marketing
  - medical diagnosis
  - treatment effectiveness analysis

Classification problem – another way …

- General task: assigning a decision class label to a set of unclassified objects described by a fixed set of attributes (features).
- Given a set of pre-classified examples, discover the classification knowledge representation,
  - to be used either as a *classifier* to classify new cases or
    to *describe* classification situations in data.
  - *Supervised* learning: classes are known for the examples used to build the classifier.
  - can be a set of rules, a decision tree, a neural network, etc.

Supervised classification - Classifiers

Learn a method for predicting the instance class from pre-labeled (classified) instances



Many approaches:
Regression,
Decision Trees,
Bayesian,
Neural Networks,
...

Given a set of points from classes ● ●
what is the class of new point ○?

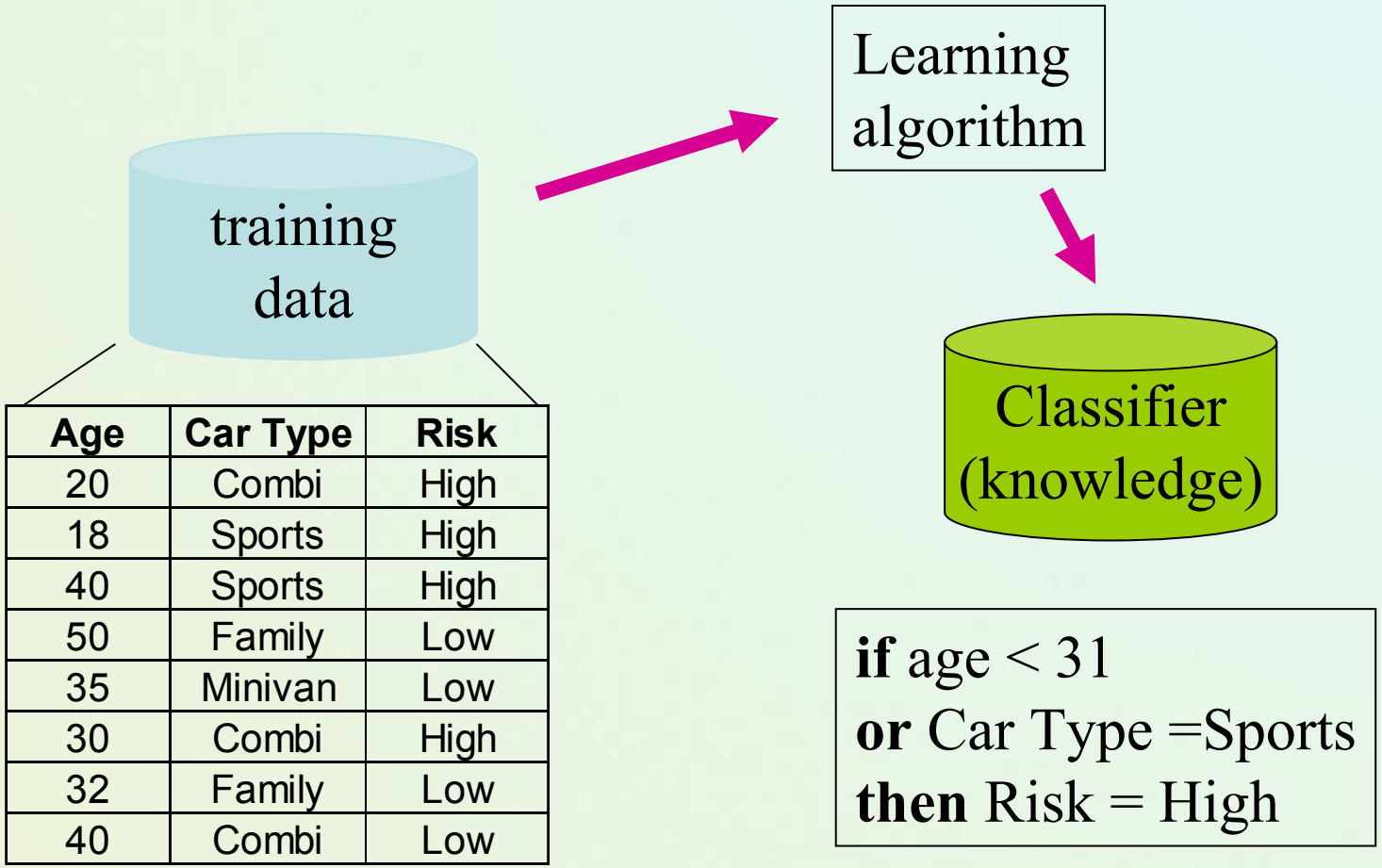
Discovering classification knowledge

Creating classifiers is a multi-step approach:

- Generating a classifier from the given learning data set,
- Evaluation on the test examples,
- Using for new examples.

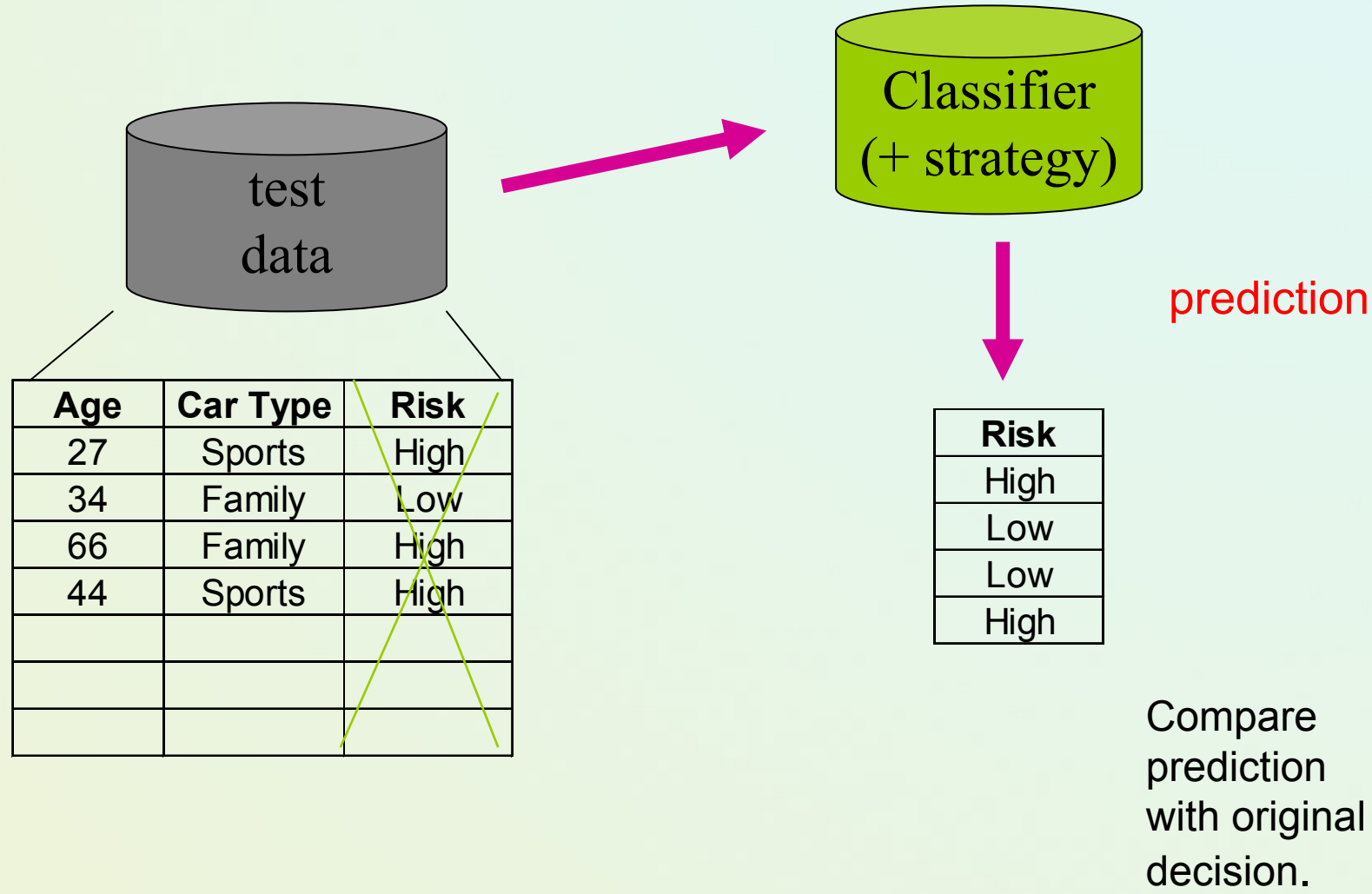
Train and test paradigm!

Discovery step

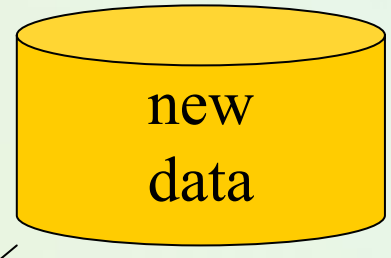


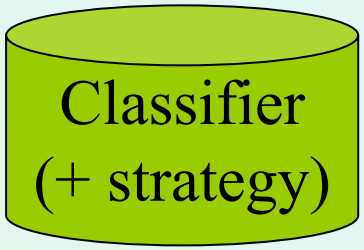
Age	Car Type	Risk
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low

Evaluation step (testing a classifier)



Predicting new data





prediction

Age	Car Type	Risk
27	Sports	?
34	Minivan	?
55	Family	?
34	Sports	?

Risk
High
Low
Low
High

Comparing Classification Methods (1)

- *Predictive* (*Classification*) *accuracy*: this refers to the ability of the model to correctly predict the class label of new or previously unseen data:
  - accuracy = % of testing set examples correctly classified by the classifier
- *Speed*: this refers to the computation costs involved in generating and using the model
- *Robustness*: this is the ability of the model to make correct predictions given noisy data or data with missing values

Comparing Classification Methods (2)

- *Scalability*: this refers to the ability to construct the model efficiently given large amount of data
- *Interpretability*: this refers to the level of understanding and insight that is provided by the model
- *Simplicity*:
  - decision tree size
  - rule compactness
- Domain-dependent quality indicators

Predictive accuracy / error

- General view (statistical learning point of view):
- Lack of generalization – prediction risk:

$$R(f) = E_{xy} L(y, f(x))$$

- where $L(y, \hat{y})$ is a loss or cost of predicting value $\hat{y}$ when the actual value is y and *E* is expected value over the joint distribution of all (*x,y*) for data to be predicted.
- Simple classification $\rightarrow$ zero-one loss function

$$L(y, \hat{y}) = \begin{cases} 0 & if\ y = f(y) \\ 1 & if\ y \neq f(y) \end{cases}$$

Evaluating classifiers – more practical …

Predictive (classification) accuracy

- Use testing examples, which do not belong to the learning set
  - N_t – number of testing examples
  - N_c – number of correctly classified testing examples
- Classification accuracy:

$$\eta = \frac{N_c}{N_t}$$

•Other options:

•analysis of confusion matrix

•misclassification costs

•Sensitivity and Specificity measures / ROC curve

Confusion matrix

	Predicted		
Original classes	K_1	K_2	K_3
K_1	50	0	0
K_2	0	48	2
K_3	0	4	46

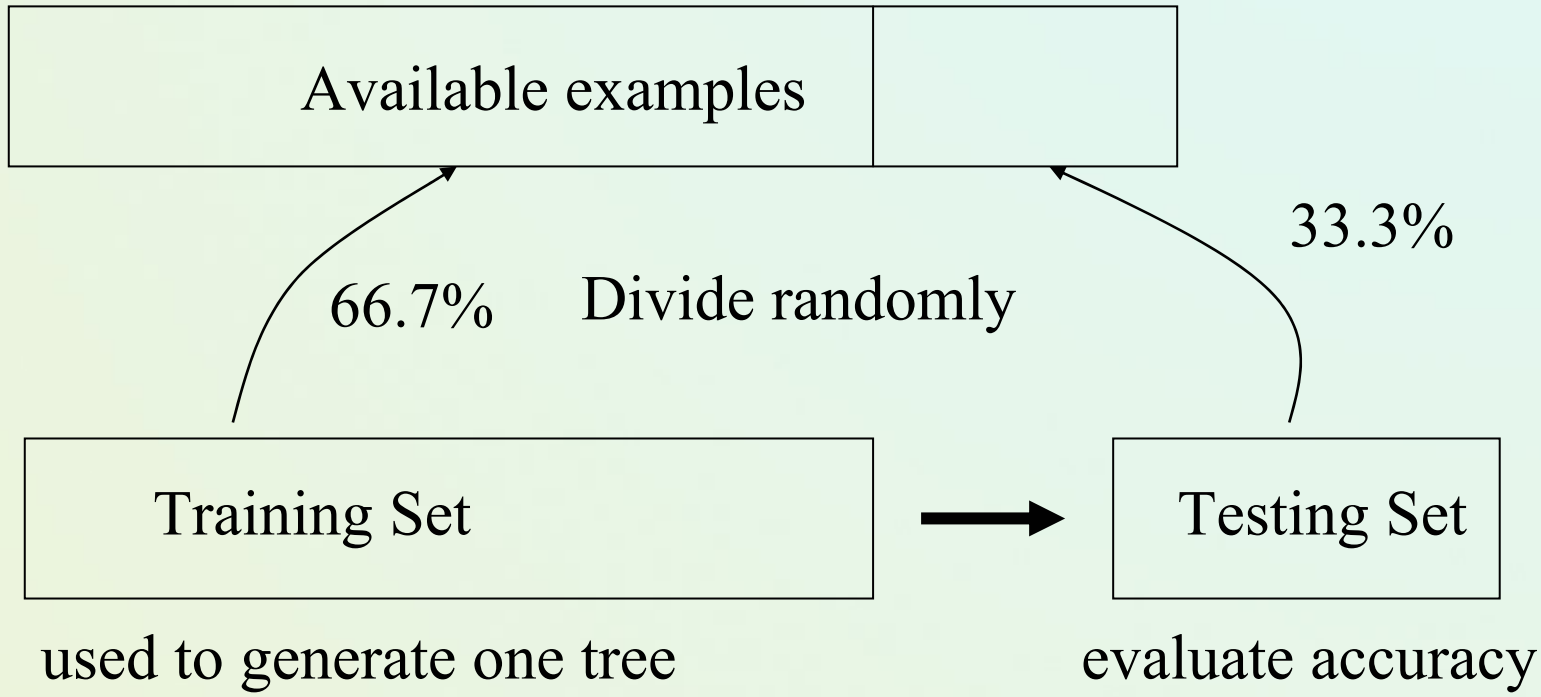
- Cost matrix / errors

$$\mathrm{C}(\varepsilon) = \sum_{i=1}^{r} \sum_{j=1}^{r} n_{ij} \cdot c_{ij}$$

Experimental estimation of classification accuracy

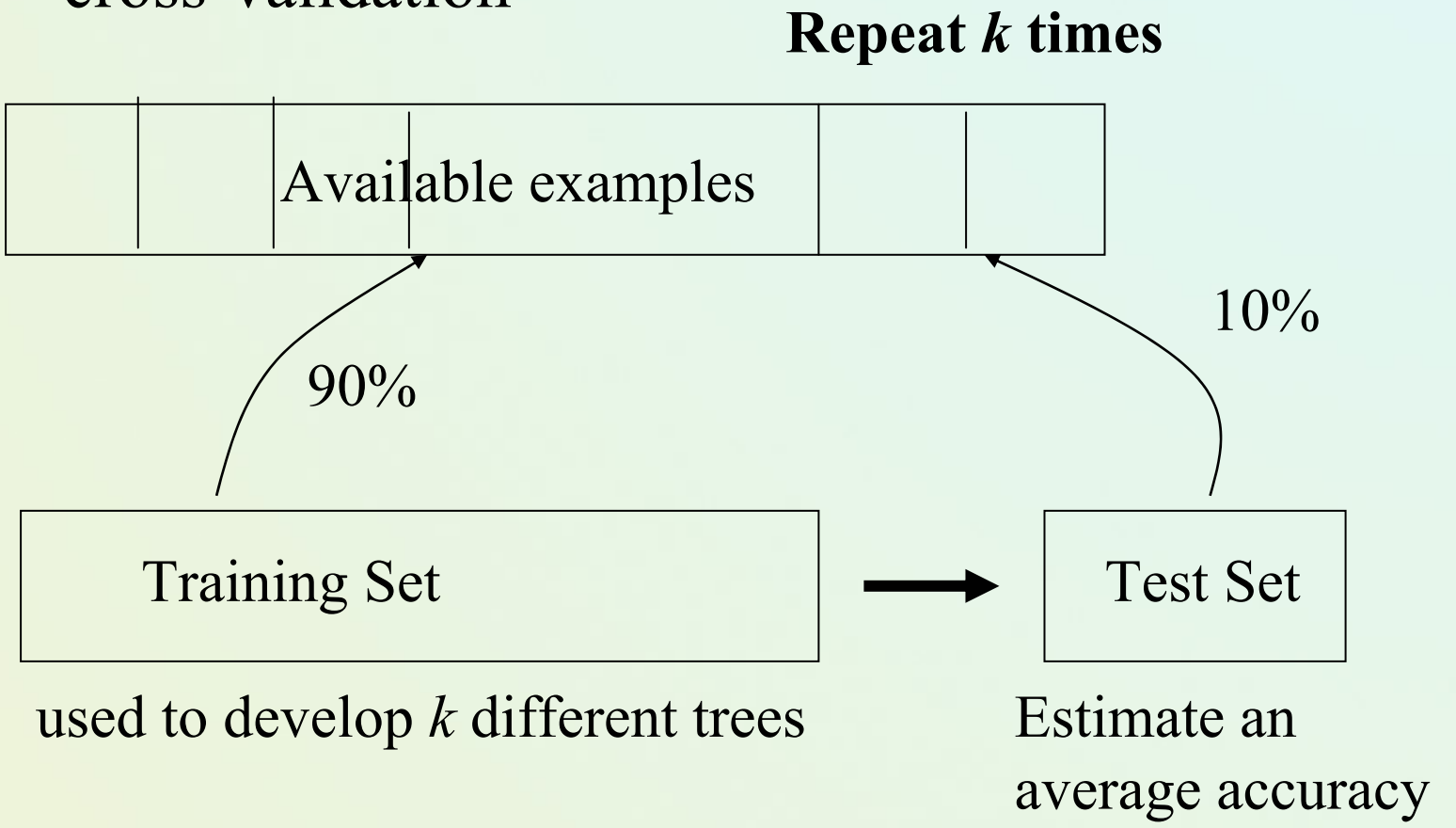
- Random partition: Training-and-testing
  - use two independent data sets, e.g., training set (2/3), test set(1/3); random sampling
- *k*-fold cross-validation
  - randomly divide the data set into *k* subsamples
  - use *k-1* subsamples as training data and one sub-sample as test data --- repeat *k* times
- leave-one-out for small size data

Evaluation tree classifiers – large dataset



Testing – small dataset

* cross-validation



More on 10 fold cross-validation

- Standard method for evaluation: stratified ten-fold cross-validation
- Why ten? Extensive experiments have shown that this is the best choice to get an accurate estimate (since CART)
- Stratification reduces the estimate's variance!
- Even better: repeated stratified cross-validation
  - E.g. ten-fold cross-validation is repeated ten times and results are averaged (reduces the variance)!

Leave-One-Out cross-validation

- Leave-One-Out:
  a particular form of cross-validation:
  - Set number of folds to number of training instances
  - i.e., for *n* training instances, build classifier *n* times but from *n* -1 training examples …
- Makes best use of the data.
- Involves no random subsampling.
- Very computationally expensive!

*The bootstrap

- CV uses sampling *without replacement*
  - The same instance, once selected, can not be selected again for a particular training/test set
- The *bootstrap* uses sampling *with replacement* to form the training set
  - Sample a dataset of *n* instances *n* times *with replacement* to form a new dataset of *n* instances
  - Use this data as the training set
  - Use the instances from the original data set that don't occur in the new training set for testing

*The 0.632 bootstrap

- Also called the *0.632 bootstrap*
  - A particular instance has a probability of 1–1/*n* of *not* being picked
  - Thus its probability of ending up in the test data is:

$$\left(1-\frac{1}{n}\right)^{n} \approx e^{-1} = 0.368$$

  - This means the training data will contain approximately 63.2% of the instances

Comparing data mining schemes

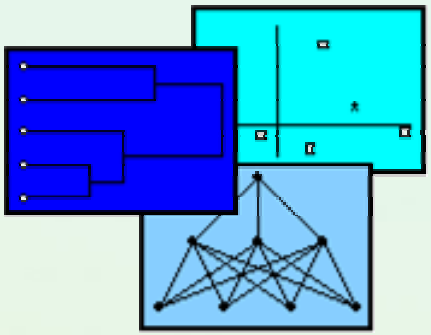
- Frequent situation: we want to know which one of two learning schemes performs better.
- Note: this is domain dependent!
- Obvious way: compare 10-fold CV estimates.
- Problem: variance in estimate.
- Variance can be reduced using repeated CV.
- However, we still don’t know whether the results are reliable.

Significance tests

- Significance tests tell us how confident we can be that there really is a difference.
- *Null hypothesis*: there is no “real” difference.
- *Alternative hypothesis*: there is a difference.
- A significance test measures how much evidence there is in favor of rejecting the null hypothesis.
- Let’s say we are using 10 times 10-fold CV.
- Then we want to know whether the two means of the 10 CV estimates are significantly different.
  - *Student’s paired t-test* tells us whether the means of two samples are significantly different.

Approaches to learn classifiers

- Decision Trees
- Rule Approaches
- Bayesian Classifiers
- Neural Networks
- Discriminant Analysis
- k-nearest neighbor classifiers
- Artificial Neural Networks
- Support Vector Machines
- Genetic Classifiers



Issues Regarding Classification and Prediction (1): Data Preparation

- Data cleaning
  - Preprocess data in order to reduce noise and handle missing values
- Relevance analysis (feature selection)
  - Remove the irrelevant or redundant attributes
- Data transformation
  - Generalize and/or normalize data

Any questions, remarks?